

Investigate Business Hotel using Data Visualization



Created by:


Muhammad Rizki Mardanu Hilman

Your Email : rizkimardanu9@gmail.com

linkedIn : linkedin.com/in/rizkimardanu

Stay curious, keep exploring. Graduated from Bandung Institute of Technology. Passionate about Data Science and data analysis. Problem finding and process-oriented data analyst with in-depth knowledge of machine learning, big data capture, analyzing and processing data using BI tools.

Had experience handling big data using SQL, Google Data Studio, and Python on project programming as a Bootcamp Candidate. lovely experience executing some BI tools; jupyter notebook, spyder to Processing Data, Statistics, Data Visualization, and Machine Learning. This moment helped me gain confidence and faith in data scientists with result-based thinking is a must within an organization. Remember something that you can't measure, you can't control. In short, a well-performed Data Scientist with result-based thinking, strong attention to detail, good execution using tools, and an initiative person will help to provide indicators and information to strengthen, sustainables and organize your organization.

A faded, light grey background image of a city skyline with various skyscrapers and buildings, providing a professional and modern aesthetic.

"It is crucial for a company to consistently analyze its business performance. In this instance, we will delve deeper into the hospitality industry. Our focus is to understand how our customers behave when making hotel reservations and its correlation with the hotel reservation cancellation rate. The insights we discover will be presented in the form of data visualization to make them more easily comprehensible and persuasive."

Data provide by Rakamin - hotel_bookings data.csv

Data Description:

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has been removed from the data.

This dataset contains 119,390 samples.

Contains 29 features :

Feature Name	Feature Description
hotel	Type of hotel
is_canceled	Cancellation status, whether the booking was cancelled (1) or not (0)
lead_time	Lead time
arrival_date_year	Year of arrival date
arrival_date_month	Month of arrival date
arrival_date_week_number	Week number of year for arrival date
arrival_date_day_of_month	Day of arrival date
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed
stays_in_weekdays_nights	Number of weekday nights (Monday to Friday) the guest stayed
adults	Number of adults
children	Number of children
babies	Number of babies
meal	Type of meal booked
city	City of origin
market_segment	Market segment designation
distribution_channel	Booking distribution channel

is_repeated_guest	Repeated guest status, whether the booking name was a returning guest (1) or a new guest (0)
previous_cancellations	Number of previous bookings that were cancelled by the customer prior to the current booking
previous_bookings_not_canceled	Number of previous bookings that were not cancelled (confirmed) by the customer prior to the current booking
booking_changes	Number of booking changes
deposit_type	Deposit type
agent	ID of the travel agency that made the booking
company	ID of the company that made the booking
days_in_waiting_list	Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type	Type of booking
adr	Average daily rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
required_car_parking_spaces	Number of car parking spaces required by the customer
total_of_special_requests	Total of special requests made by the customer
reservation_status	Reservation last status

● Programming Languages : Python



● Data Preprocessing & Cleaning Library : Pandas & Numpy



● Data Visualization Library : Matplotlib & Seaborn



Basic Datasets Information

- The dataset consists of **29 columns** and **119,390 rows** of data.
- There are 3 types of data: **int64**, **object**, **float64**.
- There are some **missing values in the following columns**:
 - **company** with a total of 94% null values, amounting to 112,593 rows.
 - **agent** with a total of 13% null values, amounting to 16,340 rows.
 - **city** with a total of 0.4% null values, amounting to 488 rows.
 - **children** with a total of 0.003% null values, amounting to 4 rows.

Data columns (total 29 columns):

#	Column	Non-Null Count	Dtype
0	hotel	119390 non-null	object
1	is_canceled	119390 non-null	int64
2	lead_time	119390 non-null	int64
3	arrival_date_year	119390 non-null	int64
4	arrival_date_month	119390 non-null	object
5	arrival_date_week_number	119390 non-null	int64
6	arrival_date_day_of_month	119390 non-null	int64
7	stays_in_weekend_nights	119390 non-null	int64
8	stays_in_weekdays_nights	119390 non-null	int64
9	adults	119390 non-null	int64
10	children	119386 non-null	float64
11	babies	119390 non-null	int64
12	meal	119390 non-null	object
13	city	118902 non-null	object
14	market_segment	119390 non-null	object
15	distribution_channel	119390 non-null	object
16	is_repeated_guest	119390 non-null	int64
17	previous_cancellations	119390 non-null	int64
18	previous_bookings_not_canceled	119390 non-null	int64
19	booking_changes	119390 non-null	int64
20	deposit_type	119390 non-null	object
21	agent	103050 non-null	float64
22	company	6797 non-null	float64
23	days_in_waiting_list	119390 non-null	int64
24	customer_type	119390 non-null	object
25	adr	119390 non-null	float64
26	required_car_parking_spaces	119390 non-null	int64
27	total_of_special_requests	119390 non-null	int64
28	reservation_status	119390 non-null	object

dtypes: float64(4), int64(16), object(9)

For the detail of my codes, you can see [here](#)

Handling Duplicate Rows
& Missing Values



Data Types Correction



Handling Invalid Values



Drop Unnecessary Data



Exploring Business
Insights

Handling Duplicate Rows

Number of duplicate rows : 33261

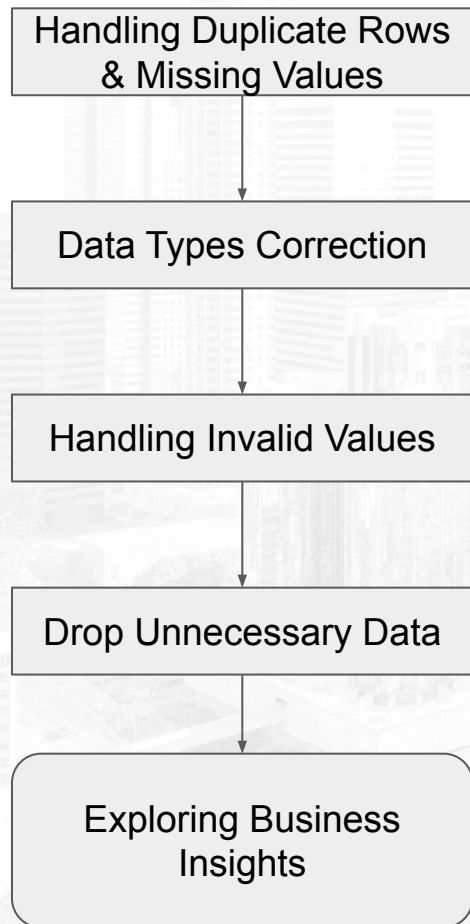
This dataset contains numerous duplicates, amounting to 33,261 rows. Nevertheless, during this process, we will refrain from eliminating the duplicate rows, as we presume that their presence is attributable to the absence of customer IDs in the data.

Handling Missing Value

	variable	dtype	count	unique	missing
0	company	float64	119390	352	112593
1	agent	float64	119390	333	16340
2	city	object	119390	177	488
3	children	float64	119390	5	4

- A company column has 94% null values, totaling 112,593 rows. We're filling zero values in the company column since there is no involvement of any company.
- An agent column has 13% null values, totaling 16,340 rows. We're filling zero values in the agent column since there is no involvement of any agent.
- A city column has 0.4% null values, totaling 488 rows. We're filling 'unknown' for entries where the city information is unavailable.
- The children column has 0.003% null values, totaling 4 rows. We're filling zero values for the children column, as it is likely that the customers have no children.

For the detail of my codes, you can see [here](#)



Data Types Correction

Change the data type of **float64** which had null before, **children**, **agent**, and **company** to **int64**

Handling Invalid Values

Replacing incorrect values in the meal column. These values will be substituted with 'No Meal' as it is assumed that 'Undefined' signifies customers who have not ordered any meals.

Drop Unnecessary Data

So, we will filter certain criteria, which are as follows:

1. Total Guest (Number of Customers / Guests) ≤ 0 , or when there are no guests present.
2. Total duration of the night ≤ 0 , or when data for the duration of the stay is not available.
3. If there is a single data entry for adr (Average Daily Rate), it could be attributed to a data calculation error. As there is only one row, it will be removed to prevent errors in the analysis.

For the detail of my codes, you can see [here](#)

Monthly Hotel Booking Analysis Based on Hotel Type

In the hospitality sector, customer booking patterns within the hotel industry are of paramount importance since they have a direct influence on the company's income. It is imperative to conduct an in-depth analysis of customer booking behaviors when reserving hotel accommodations. For instance, we can discern which categories of hotels are favored by customers and establish connections with the seasonal factors influencing hotel reservations.

Hence, the objective of this endeavor is to contrast the monthly count of hotel reservations according to hotel categories and evaluate the periods of rising or declining booking activity across various months and seasons.

Actions undertaken:

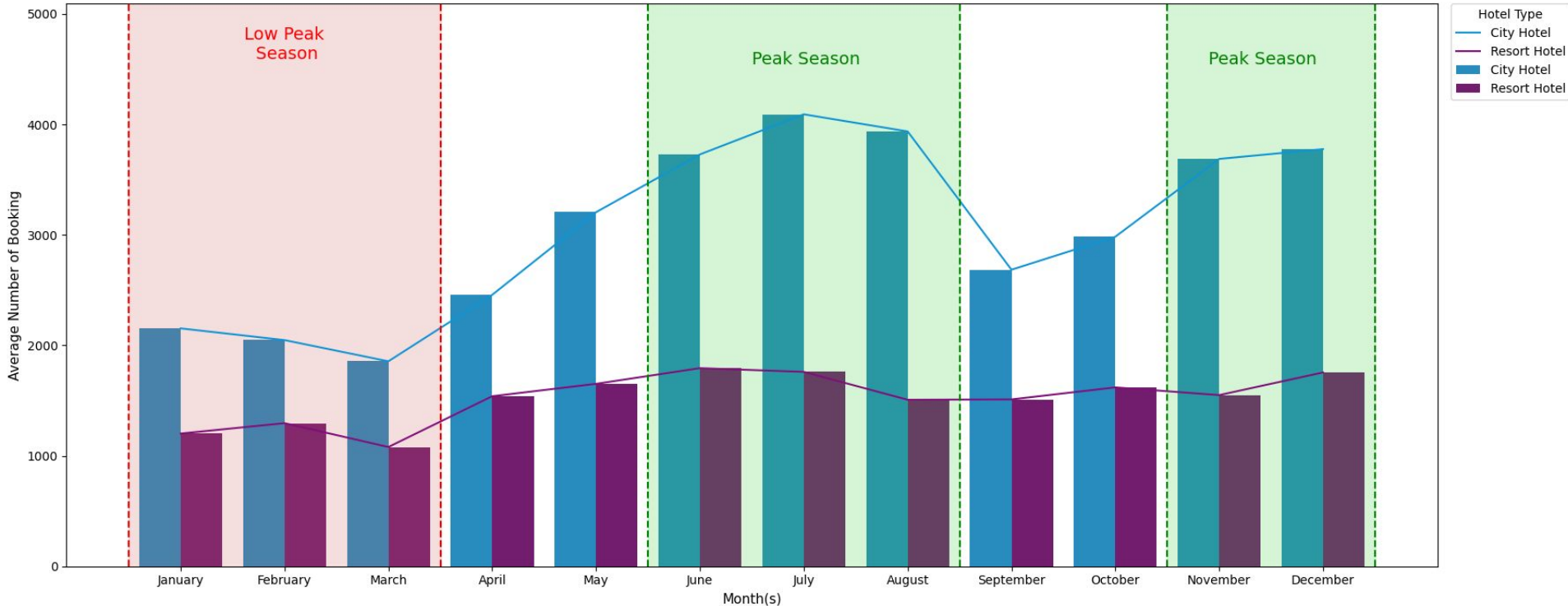
1. Generate a consolidated table illustrating the monthly comparison of hotel bookings according to the various hotel types, with particular focus on the arrival year data.
2. Standardize the data, giving special consideration to the information pertaining to the months of September and October.
3. Arrange the data in ascending order based on the months, ensuring accurate spelling of the month names to facilitate visualization.
4. Develop a line graph to depict variations in the fluctuations of hotel bookings on a monthly basis, categorized by hotel types.
5. Explanation

For the detail of my codes, you can see [here](#)

Monthly Hotel Booking Analysis Based on Hotel Type

Average Number of Hotel Bookings per Month Based on Hotel Type

Hotel bookings show a significant increase during the peak season and high season holidays. In June - August, this is primarily due to school holidays and Eid al-Fitr holidays in Indonesia. In November and December, the increase in bookings is attributed to the New Year holidays.



For the detail of my codes, you can see [here](#)

Interpretation:

1. Both hotels are experiencing a similar upward trend in the average number of hotel bookings. However, the highest peak occurs at the City Hotel.
2. Hotel bookings exhibit a significant increase during the high season and holiday periods. The high season spans from June to August for both the City Hotel and the Resort Hotel, primarily driven by school holidays and the extended Eid al-Fitr holiday season in Indonesia.
3. Furthermore, there is another high season in November and December, albeit of shorter duration compared to the high season in June and July, likely due to New Year holidays and the conclusion of annual leave allotments. To optimize hotel room reservations during the low season, the hotel can implement New Year promotions to attract a higher number of visitors.
4. The Low Season extends throughout January to March and also in August to September, with a notable decrease in bookings, especially at the City Hotel. This decline can be attributed to the commencement of the new school and office seasons, as students and professionals shift their focus towards academic and vocational activities. During the low season, the hotel may consider offering discounts or vouchers to entice customers to continue patronizing the establishment.

For the detail of my codes, you can see [here](#)

This analysis centers on the relationship between the length of stay and the hotel booking cancellation rate. According to the data, approximately 19% of hotel reservations made online are canceled before guests arrive.

These cancellations have the potential to diminish room availability and can impact the hotel's revenue, as each vacant room can pose a financial burden for that day. Furthermore, if the hotel utilizes an Online Travel Agency (OTA), these cancellation rates may have implications for the hotel's search rankings.

Steps taken

1. Pay attention to the "stay duration" column, which is obtained by summing the weekdays and weekend nights. Next, we will examine the data distribution to simplify the grouping process accordingly.
2. Group the values of the new column obtained in the previous step to make it more significant, taking into account the data distribution for meaningful categorization.
3. Create an aggregate table that compares the number of canceled hotel bookings based on the duration of stay for each hotel type, focusing on the proportion of canceled bookings.
4. Create a bar plot to display the cancellation ratio of bookings based on the duration of stay for each hotel type, emphasizing the proportion of canceled bookings.
5. Interpretation

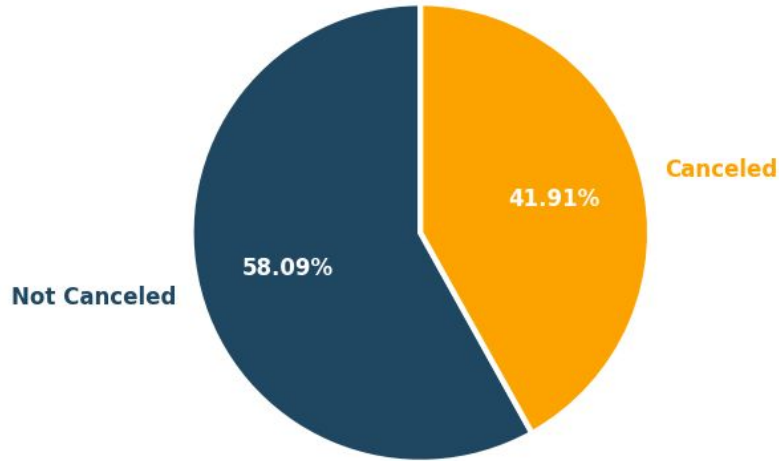
For the detail of my codes, you can see [here](#)

Impact Analysis of Stay Duration on Hotel Bookings Cancellation Rates

Rasio Total Pembatalan Pemesanan Hotel City

	is_canceled	total
0	Not Canceled	45833
1	Canceled	33066

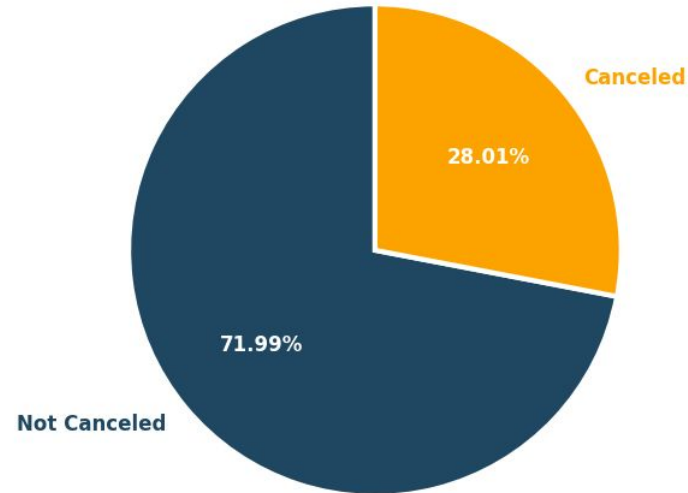
Rasio Total Pembatalan Pemesanan Hotel City



Rasio Total Pembatalan Pemesanan Hotel Resort

	is_canceled	total
0	Not Canceled	28555
1	Canceled	11110

Rasio Total Pembatalan Pemesanan Hotel Resort



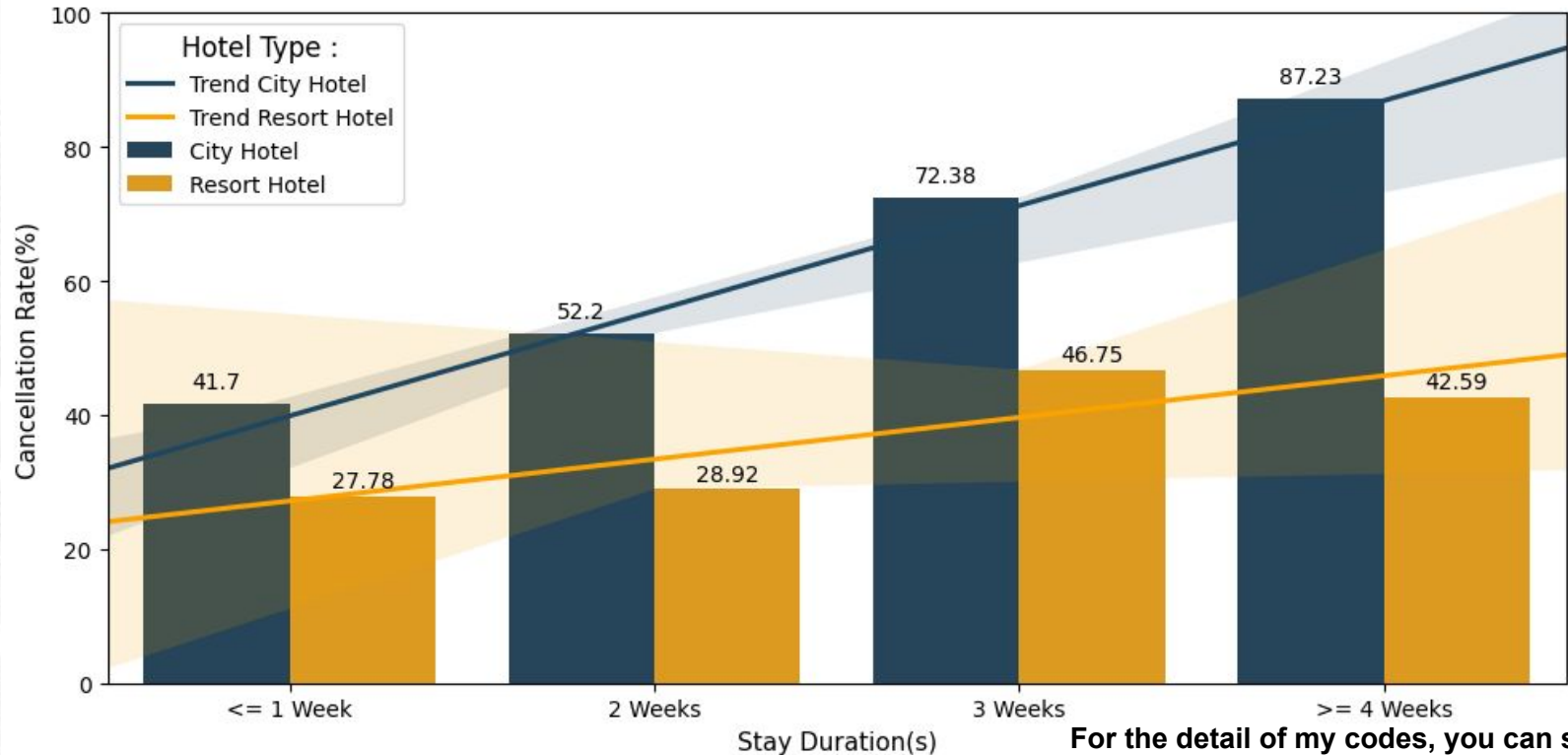
For the detail of my codes, you can see [here](#)

Cancellation Rate Trend Based on Stay Duration and Hotel Types

Longer stays have higher cancellation rates.

City Hotel: Highest cancellation rate at four weeks stay duration (87.23%).

Resort Hotel: Highest cancellation rate at three weeks stay duration (46.75%).



Interpretation :

1. Based on the length of the stay, it is evident that **the hotel booking cancellation rates tend to increase as the duration of the stay becomes longer, both at the City Hotel and the Resort Hotel**. Specifically, the City Hotel experiences a more pronounced rise in cancellation rates, dipping below 50% for one-week stays. Conversely, at the Resort Hotel, cancellations tend to be lower for stays exceeding four weeks. Essentially, both exhibit a positive trend, where the longer the stay duration, the higher the likelihood of reservations being canceled.
2. The reasons behind the increase in cancellations with longer durations of stay require further evaluation and analysis by the company. Several factors may come into play, including customer dissatisfaction. Guests with extended stays may have higher expectations and a desire for better services. If they feel dissatisfied with the hotel's service, they may decide to cancel their reservation and seek a hotel that better meets their expectations. Additionally, higher costs can also be a consideration. Expenses may exceed the customer's initial estimate or budget, prompting them to look for more affordable alternatives.
3. Business recommendations provided: Implement a deposit or partial payment policy when customers make a booking to encourage more prudent decision-making when considering cancellations. Furthermore, a policy could be established that restricts the cancellation of bookings when the current date is in close proximity to the booking date (for example, no cancellations allowed if it is within 7 days of the stay).

For the detail of my codes, you can see [here](#)

The purpose of this investigation into business insights is to assess the relationship between the lead time for hotel reservations and the frequency of hotel booking cancellations. Within the hotel industry, patrons are typically given the option to book accommodations prior to their scheduled arrival, with lead times spanning from a few days to several months. The objective is to investigate if the interval between making a hotel reservation and the actual check-in date has an impact on the rate of hotel booking cancellations.

Steps taken

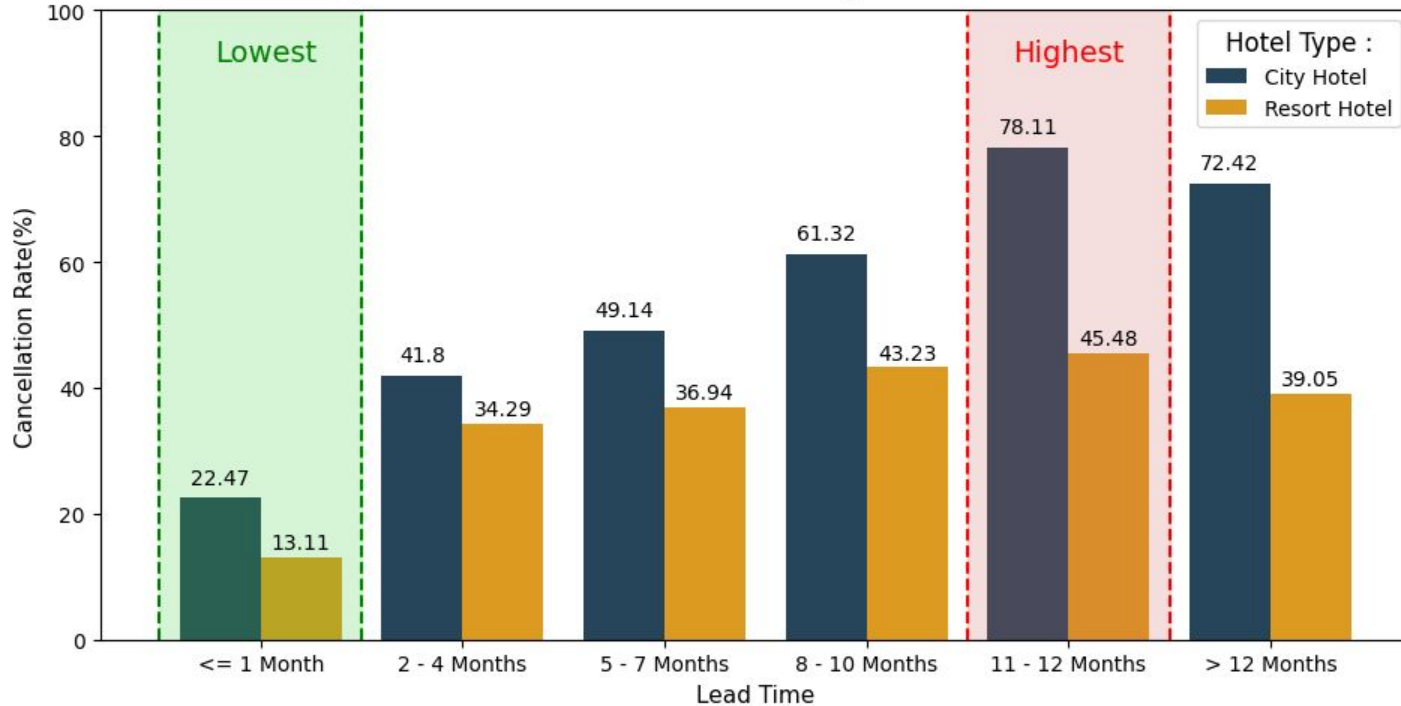
1. Create a new column that categorizes the booking lead time by creating intervals for the categorization.
2. Create an aggregated table comparing canceled hotel bookings based on the booking lead time for each hotel type, focusing on the proportion of canceled bookings.
3. Create a plot illustrating the cancellation ratio of bookings based on the booking lead time for each hotel type, using an appropriate plot type.
4. Interpretation

Impact Analysis of Lead Time on Hotel Bookings Cancellation Rate

Cancellation Rate Trend Based on Stay Duration and Hotel Types

Longer lead times increase order cancellation probability.

Both hotel types have the lowest cancellation rate in ≤ 1 month
and have the highest cancellation rate for bookings made with 11-12 months lead time



For the detail of my codes, you can see [here](#)

Interpretation :

1. In general, a longer lead time is associated with a greater likelihood of order cancellation. Lead time refers to the number of days between the booking entry into the Property Management System (PMS) and the arrival date, and longer lead times are linked to higher cancellation rates.
1. Both hotel categories exhibit their lowest cancellation rates for lead times of one month or less, with the City Hotel at 22.47% and the Resort Hotel at 13.11%.
1. For bookings made with a lead time of 11-12 months, both hotel types experience their highest cancellation rates, with the City Hotel at 77.41% and the Resort Hotel at 43.5%.
1. Resort and City Hotels both encounter their highest cancellation rates when the lead time is approximately one year. This could be attributed to customers' vacation plans being canceled or them forgetting about their hotel reservations when the lead time is excessively long. To mitigate cancellations, hotels can send reminders to customers and implement a stringent cancellation policy for all reservations to minimize such occurrences.

Summary:

1. Both hotels are exhibiting a similar upward trajectory in their reservation patterns, with the City Hotel reaching the highest peak. Hotel reservations experience a substantial surge during peak and high season holidays, specifically in June to August and November to December. Notably, the City Hotel witnesses a significant drop in reservations throughout January to March and August to September. To enhance their reservation rates during off-peak seasons, implementing promotions and discounts can prove to be effective strategies for both establishments.
1. The City Hotel records the highest rate of cancellations, showing a noteworthy upward trend. There exists a positive correlation between the duration of stay and the rate of cancellations in both hotels. The City Hotel experiences its highest cancellation rate for stays exceeding 4 weeks, while the Resort Hotel's peak rate occurs for stays of 3 weeks. Enforcing more stringent cancellation policies and offering exclusive promotions can be instrumental in mitigating the occurrence of cancellations.
1. Longer lead times are linked to higher cancellation rates for both the City and Resort Hotels. The most favorable cancellation rates are observed when the lead time is less than or equal to 1 month, whereas the highest rates are associated with lead times of 11-12 months. Notably, both Resort and City Hotels experience their peak cancellation rate at approximately a 1-year lead time. The implementation of reminders and rigorous cancellation policies can play a pivotal role in decreasing cancellations and enhancing overall booking efficiency.